CLAIMS

Having thus described my invention, what I claim as new and desire to secure by Letters Patent is as follows:

1. A method of classifying a source publishing a document on a portion of a network, comprising steps of:

electronically receiving a document;

based on the document, determining a source which published the document; and

assigning a code to said document based on whether data associated with the document published by the source matches with data contained in a database.

2. The method according to claim 1, wherein said portion of said network comprises a graphical multimedia portion of said network, said source comprises a Web site publishing a home page, and said network comprises the Internet.

3. The method according to claim 2, wherein said graphical multimedia portion of said network comprises the World-Wide Web (WWW) and said document comprises a Web document,

wherein said step of assigning a code includes determining that the Web site comprises a first entity when there is a match of the Web site with said data, and determining that the Web site comprises a second entity when there is no match of the Web site with said data.

4. The method according to claim 1, wherein said step of determining a source includes:

extracting a domain name from a predetermined uniform resources locator (URL) database;

querying a database for storing registered domain names; and

merging an address database with predetermined data.

5. The method according to claim 4, wherein said predetermined data comprises Yellow Pages data,

wherein said step of determining further comprises:

characterizing uniform resource locators (URLs) by their Internet Protocol (IP) addresses including identifying a plurality of attributes based on the IP addresses of new URLs, a new URL being retrieved and parsed into a domain name and directory path portions, and

determining, based on said domain name, whether a selected URL is hosted on one of a true server and a shared server.

6. The method according to claim 5, said step of determining further comprising:

for a shared server, determining a root path by searching for the given domain name in a new URL database and identifying common directory paths,

wherein, when no match is present, the URL is processed subsequently at a later iteration, and, when a match is present, the root path is set to a matching path.

7. The method according to claim 6, wherein said step of assigning a code comprises:

automatically identifying a business associated with the source publishing said document, said business being hosted on a Service Provider (SP) Web server.

8. The method according to claim 7, wherein said step of assigning a code further comprises:

receiving a URL based on said determining step; and a URL determining step for determining whether said URL comprises one of a root URL and a leaf URL.

9. The method according to claim 8, wherein said root URL comprises an entry point for a home page on the World-Wide Web, and a leaf URL comprises a link below a root URL,

wherein said URL determining step comprises:

parsing said URL into a domain name component and a directory path component;

analyzing the domain name in said domain name component to determine whether it is associated with an SP;

when the domain name is not associated with an SP, checking the directory path component to judge whether a directory path is missing, a missing directory path indicating a root URL;

when the domain name is associated with an SP, checking whether a directory path does not exist to thereby determine that said domain name comprises a root URL, and when a directory path exists, then comparing the path to known SP Client Directory paths.

10. The method according to claim 9, further comprising:

when said URL is determined to be a root URL, analyzing a home page associated with said root URL automatically to extract home page data contained therein and assigning the home page data to the Root URL being analyzed.

11. The method according to claim 10, further comprising:

comparing said home page data with data in a predetermined business organizations database,

wherein, when there is a match, said code is assigned to the corresponding root URL, and, when no match is found, said URL is identified for subsequent analysis of lower-level hyperlinks during a next iteration of said method.

12. The method according to claim 11, wherein when no match is found at any level, said home page is identified as a personal page.

13. A method of automatically assigning a document a code for distinguishing a first-type page from a second-type page, comprising steps of:

electronically receiving a document;

based on the document, determining a source which published the document; and

assigning a code to said document based on whether the source matches with data contained in a database.

14. A search engine for use on a network for distinguishing between business web pages and personal web pages, comprising:

means for parsing the content of a hyper-text markup language (HTML) at a web address and searching for criteria contained therein;

means for analyzing a uniform resources locator (URL) of the web address to determine characteristics thereof of a web page at the web address;

means for determining whether said criteria match with data contained in a database; and

means for cross-referencing a match, determined by said determining means, to a second database, to classify a source which published the web page.

15. A search engine according to claim 14, wherein said criteria include at least one of an address, a telephone numbers, a facsimile number, a contact and a key-word contained in said HTML, and

wherein the characteristics of said web page include a geographical location and a web host computer.

16. A search engine according to claim 14, wherein said database includes a Business Semantic Terminology database having information related to business categories in a Yellow Pages directory.

17. A search engine according to claim 14, wherein said second database includes a Yellow Pages database.

18. A search engine according to claim 14, wherein said web page comprises hyperlinks, and said means for parsing comprises an indexer robot for traversing said hyperlinks in said web page and a web index database,

said indexer robot for indexing a content of said web page into said web index database.

19. A search engine according to claim 14, wherein said means for analyzing comprises:

means for determining whether said URL comprises one of a root URL and a leaf URL.

20. A search engine according to claim 19, wherein said root URL comprises an entry point for the web page on the World-Wide Web, and a leaf URL comprises a link below a root URL, said search engine further comprising:

means for parsing said URL into a domain name component and a directory path component;

means for analyzing the domain name in said domain name component to determine whether it is associated with an SP;

means for checking the directory path component to judge whether a directory path is missing, when the domain name is not associated with a service provider (SP), a missing directory path indicating a root URL, and for checking whether a directory path does not exist to thereby determine that said domain name comprises a root URL, when the domain name is associated with an SP;

means for comparing the path to known SP Client Directory paths, when a directory path exists;

means for analyzing a home page associated with said root URL, when said URL is determined to be a root URL, thereby automatically to extract home page data contained therein; and

means for assigning the home page data to the Root URL being analyzed.